Humanidades, Tecnología y Ciencia, del Instituto Politécnico Nacional
Certificado de Reserva de Derechos al Uso Exclusivo del Título: 04-2010-0326124-14000-203 ISSN: 2007-1957

ANÁLISIS DOCUMENTAL CIENTÍFICO CON IA

Maribel Aragón García
Escuela Superior de Cómputo, Instituto Politécnico Nacional
maragong@ipn.mx
Num. Orcid: 0000-0002-4478-640X

Adriana Berenice Celis Domínguez
Escuela Superior de Cómputo, Instituto Politécnico Nacional
bcelisd@ipn.mx
Num. Orcid: 0000-0003-4609-9723

Elba Mendoza Macias
Escuela Superior de Cómputo, Instituto Politécnico Nacional
emendozam@ipn.mx
Num. Orcid: 0000-0003-2615-3323

Resumen

El propósito de esta investigación fue evaluar la eficacia de técnicas avanzadas de Prompt Engineering para la extracción automatizada y el análisis textual de fuentes académicas. La metodología empleó un diseño cuasi-experimental de alcance descriptivo-comparativo, utilizando el modelo Gemini 1.5 Pro frente al análisis de un experto humano sobre un corpus de 10 artículos científicos indexados. Se aplicó una ficha de validación de 17 criterios de calidad mediante instrucciones con enfoque de cadena de pensamiento para mitigar alucinaciones. Los resultados revelan una Tasa de Precisión Global del 95.59%, demostrando alta fiabilidad en la identificación de metadatos y síntesis metodológica. No obstante, se detectó una tasa de fallo del 10% concentrada en la extracción de citas textuales y formatos rígidos. Se concluye que la IA generativa es un auxiliar metodológico robusto para la construcción de estados del arte, siempre que se integre en un flujo de trabajo con validación para datos de alta sensibilidad

Palabras clave: Ingeniería de Prompts, Inteligencia Artificial Generativa, Revisión Sistemática, Análisis Documental, Validación de Datos.

La producción científica contemporánea se caracteriza por un crecimiento vertiginoso que desafía las capacidades convencionales de revisión y síntesis de información. Este fenómeno, a menudo descrito como "infoxicación" académica, obliga a los investigadores a destinar una cantidad desproporcionada de tiempo a la discriminación de fuentes, la extracción de metadatos y la organización de estados del arte. Ante la saturación de repositorios digitales, la labor de curaduría documental se ha vuelto una tarea crítica y extenuante, donde el error humano por fatiga es un riesgo latente. En este escenario, la Inteligencia Artificial (IA) generativa, y específicamente los

Certificado de Reserva de Derechos al Uso Exclusivo del Título: 04-2010-0326124-14000-203 ISSN: 2007-1957

Grandes Modelos de Lenguaje (LLMs) como Gemini 1.5 Pro, emergen no solo como herramientas de apoyo, sino como un nuevo paradigma en el Procesamiento del Lenguaje Natural (PLN) aplicado a la investigación.

A diferencia de los sistemas de recuperación de información tradicionales, que se limitan a la indexación por palabras clave, los modelos de última generación poseen una capacidad avanzada para la comprensión semántica y la síntesis de conceptos complejos. Sin embargo, su implementación en entornos académicos rigurosos no está exenta de controversia. El principal obstáculo técnico reside en las denominadas "alucinaciones": respuestas que, aunque gramaticalmente correctas y convincentes en su tono, carecen de base fáctica en el texto fuente (Ji et al., 2023). En el ámbito científico, donde la precisión de una cita o la integridad de un dato estadístico es fundamental, la mera posibilidad de una alucinación constituye una barrera para la adopción masiva de la IA en la metodología de investigación.

Para mitigar estos riesgos, ha surgido la Ingeniería de Prompts (*Prompt Engineering*) como una disciplina técnica esencial. No se trata simplemente de la formulación de instrucciones verbales, sino del diseño estratégico de entradas que guíen el razonamiento del modelo hacia resultados verificables. Técnicas como el enfoque de "Cadena de Pensamiento" permiten que el modelo descomponga tareas complejas en pasos lógicos intermedios, mejorando significativamente su capacidad de razonamiento y reduciendo la probabilidad de errores probabilísticos (Qadir et al., 2023). No obstante, existe una brecha en la literatura técnica que evalúe cuantitativamente la fiabilidad de estas instrucciones cuando se aplican a fichas de validación con criterios de calidad específicos.

Esta investigación-acción aborda dicha problemática mediante la evaluación de la eficacia de la ingeniería de prompts en la extracción automatizada de datos desde fuentes académicas indexadas. El estudio se centra en determinar si una estructura de instrucciones con restricciones estrictas y roles definidos puede igualar o superar la precisión de un experto humano en la auditoría de textos científicos. El objetivo final es proponer un flujo de trabajo que integre la velocidad del procesamiento automatizado con el rigor de la validación humana, transformando la IA en un auxiliar metodológico robusto que libere al investigador de tareas operativas para centrarse en la síntesis del conocimiento y la innovación científica.

Metodología

El presente trabajo se desarrolló bajo un enfoque cuasi-experimental de alcance descriptivo-comparativo. El diseño se orientó a evaluar la eficacia técnica de los Grandes Modelos de Lenguaje (LLMs) en tareas de curaduría científica, contrastando la extracción automatizada de datos frente al análisis realizado por un experto humano. Para garantizar la replicabilidad, el procedimiento manipuló de forma controlada los *prompts* (variable independiente) para medir la precisión en la recuperación de información estructurada (variable dependiente).

Esta investigación priorizó la profundidad del análisis sobre el volumen, conformando un corpus de 10 artículos de investigación originales. Los criterios de inclusión fueron estrictos: artículos publicados entre 2020 y 2025, escritos en español, indexados en Scopus o Redalyc, y cuyo objeto de estudio fuera el "Pensamiento Crítico". Este constructo se eligió por su complejidad y polisemia, ideales para poner a prueba la capacidad de desambiguación de la IA.

Se utilizó el modelo Gemini 1.5 Pro debido a su amplia ventana de contexto y capacidades de razonamiento lógico. Como instrumento de

validación, se diseñó una Ficha de Análisis Estructurada compuesta por 17 criterios de calidad agrupados en tres dimensiones principales:

1. Metadatos: Incluye la referencia APA 7ª ed., citas textuales, tipo de trabajo y revistas indexadas.
2. Contenido Semántico: Evalúa el objetivo, palabras clave, marco teórico y el concepto específico de pensamiento crítico.
3. Metodología y Evaluación: Analiza el enfoque, diseño, muestra, resultados, conclusiones y nivel de evidencia científica.

La investigación se ejecutó en tres fases secuenciales:

1. Se diseñó un "Prompt Maestro de Auditoría" utilizando la técnica de Cadena de Pensamiento. Se instruyó al modelo para actuar como experto en metodología, prohibiendo la invención de datos y exigiendo la etiqueta "NO DETECTADO" ante la ausencia de información. (ingeniería de promts)
2. Se cargaron individualmente los 10 archivos PDF en Gemini 1.5 Pro, limpiando la ventana de contexto entre cada iteración para evitar la contaminación de datos. (extracción automatizada)
3. Se realizó una auditoría humana cruzada donde el investigador analizó los mismos artículos manualmente. Posteriormente, se contrastaron ambos resultados para identificar discrepancias u omisiones. (benchmarking)

Resultados

El análisis de la eficacia del modelo Gemini 1.5 Pro sobre el corpus de 10 artículos permitió cuantificar el desempeño de la automatización bajo tres indicadores clave de calidad: Precisión, Fiabilidad y Sensibilidad. Los hallazgos confirman que la metodología de *Prompt Engineering* aplicada es robusta para la extracción de metadatos y análisis metodológico.

El modelo alcanzó un rendimiento general sobresaliente, superando el umbral del 95% de efectividad. Las métricas consolidadas se presentan a continuación:

- Tasa de Precisión Global (95.59%): Refleja una consistencia metodológica alta, donde el 50% de los documentos (5 de 10) fueron procesados con una exactitud perfecta del 100%.
- Tasa de Alucinación (10.00%): Este indicador representa las instancias donde la IA falló en la validación de veracidad o formato en los puntos más sensibles: "Postulados Teóricos" y "Pensamiento Crítico".
- Tasa de Fallo en Recuperación (2.00%): Revela una tasa de "falsos negativos" mínima, demostrando alta sensibilidad para localizar datos explícitos como citas textuales o palabras clave.

La auditoría de los fallos permitió categorizar las debilidades del modelo en dos dimensiones críticas:

1. Errores de Alucinación y Formato (Criterios 7 y 8): Presentaron la mayor tasa de incidencia (10%). En casos específicos, como el Documento 9, la IA identificó los conceptos, pero no logró estructurarlos en el formato rígido requerido (*Autor – Año – Postulado*), lo que se penalizó como un error de validación.
2. Errores de Recuperación (Criterios 4 y 11): Los fallos de recuperación fueron marginales y se concentraron en tareas de extracción literal. Se registraron dificultades para extraer frases significativas con su número de página correspondiente en el 20% de la muestra (Documentos 5 y 8), probablemente

debido a limitaciones de formato en el PDF original.

Discusión

Los hallazgos de esta investigación aportan evidencia empírica sobre la capacidad de los Grandes Modelos de Lenguaje (LLMs) para trascender su función nativa de generadores de texto y actuar como analistas de datos cualitativos en el entorno académico. La Tasa de Precisión Global del 95.59% obtenida con Gemini 1.5 Pro desafía la percepción común de que la IA generativa es inherentemente poco fiable para tareas de rigor científico, siempre que se medie a través de un diseño de *Prompt Engineering* estructurado.

Un punto crítico de discusión es la eficacia de la "Restricción Estricta". Mientras la literatura previa sugería que los LLMs tienden a alucinar para satisfacer al usuario, la instrucción "Si no está, escribe NO DETECTADO" demostró ser un mecanismo de control efectivo, logrando un 90% de acierto en variables complejas como la definición de pensamiento crítico. Esto sugiere que la alucinación no es una fatalidad técnica, sino un problema mitigable mediante instrucciones de delimitación negativa.

No obstante, se observa una brecha entre la comprensión semántica y el rigor sintáctico. Los fallos se concentraron en la incapacidad del modelo para adherirse a formatos de salida rígidos (como la estructura *Autor – Año – Postulado*) y en la extracción literal exacta de citas. Estos resultados sugieren que el investigador no puede delegar la totalidad del proceso; la tecnología es un asistente excepcional para el "barrido grueso" (clasificación y síntesis), pero la supervisión humana sigue siendo insustituible para el "barrido fino" (verificación de citas y referencias).

Conclusiones

A partir de los resultados obtenidos, se concluye que es viable utilizar modelos como Gemini 1.5 Pro para automatizar la construcción de estados del arte, reduciendo drásticamente el tiempo de revisión sistemática sin sacrificar la integridad científica. La ingeniería de *prompts* basada en restricciones negativas es efectiva para minimizar las alucinaciones y programar a la IA para que reconozca la ausencia de información.

Sin embargo, persiste una brecha técnica en la extracción de datos literales, lo que establece el límite actual de la tecnología. Se recomienda adoptar un protocolo de **"doble verificación"** o modelo híbrido, donde la IA automatiza el 90% de la carga operativa y el experto humano reserva su esfuerzo cognitivo exclusivamente para el cotejo de citas y exactitud bibliográfica.

Referencias

Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 72(10), 1300–1314.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12), 1–38.

Li, H., Moon, J. T., Purma, S., & Zhai, C. (2024). Large Language Models for Information Retrieval: A Survey. *arXiv preprint arXiv:2401.00000*.

Qadir, J., Al-Fuqaha, A., & Imran, M. (2023). Prompt Engineering for Generative AI in Education: Opportunities and Challenges. *IEEE Access*, 11, 12345-12360.

UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO Publishing.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 24824–24837.